Name of Faculty: Dr. Virendra Kumar Tiwari

Designation: Professor

Department: LNCT-MCA

Subject: 202-DBMS

Unit: V

Topic: ***Data Warehousing- terminology, definitions, characteristics.***

DATA WAREHOUSE

Definition: Data warehouse is a collection of data designed to support management decision making.

Another definition is: Data warehousing is the process, whereby, organizations extract meaning from their informational assets through the use of data warehouses.

Another definition is: A data warehouse is a subject oriented, integrated, time variant, nonvolatile collection of data used in support of management decision making processes.

The meaning of the key terms in this definition is as follows:

(i) Subject oriented: In data warehouse, data is organized and optimized according to specific subjects or areas of interest of the organization rather than simply as computer files. Examples of major subject areas include Customers, Products, Accounts, Transactions etc. It provides capability to provide answers to various queries coming from various functional areas within an organization.

(ii) Integrated: Data warehouse is integrated a single source of information for and about understanding multiple areas of interest. It provides information about a variety of subjects at one place. The input data comes from various sources in inconsistent form. Data warehouse refines the data to make it consistent and provides a unified view of overall organizational data to users. So, data warehouse is a centralized depository of data of the entire organization, which helps in better understanding of organization's operations for strategic business opportunities and increase decision making capabilities.

(iii) Non-volatile: Data warehouse contains stable information that doesn't change each time an operational process is executed. The data in the data warehouse are loaded and refreshed from operational systems, but cannot be changed by end users. New data is always added as a supplement to database, rather than a replacement.

(iv) Time-variant: The data in Data warehouse is only accurate and valid at some point in time or over some time interval. Data contains a time-dimension so that they may be used as a historical record of business' i.e., sales statistics of previous week.

(v) Accessible: The primary purpose of a data warehouse is to provide readily accessible information to end users.

A number of separate technologies have come together to make Data warehousing possible to implement. However, it may be deployed physically, the data warehouse may be viewed as a single, consistent state of information with appropriate tools to provide valuable information about a business.

1. Distinctive Characteristics of Data Warehouses

The data warehouses have many characteristics that make them different from others.

- It typically integrates several resources e.g., sales databases from various regions/states/ years.
- It requires more historical data than generally maintained in operational databases.
- It must be optimized for access to very large amounts of data.
- It is mostly read-accessed and rarely write-accessed.
- Data may be coarser grained than in operational databases.
- Data warehouses are maintained separately from operational data.
- It is based on client-server architecture.
- It provides multi-user support.
- It is capable of handling dynamic sparse matrices.
- It provides multidimensional conceptual view.

- It supports unrestricted cross-dimensional operations.
- It maintains transparency.
- It provides consistent and flexible reporting performance.
- Its having unlimited dimensions and aggregation levels.

2. *Difference between Database and Data Warehouse*

There are many differences in database and data warehouse. These differences are in the organization and data stored in both. The various differences are as follows:

S. No.	Database	Data Warehouse
1.	A data base is a collection of related data. The database system is a collection of database and DBMS.	A data warehouses is a collection of information as well as a supporting system.
2.	The data bases maintain a balance between efficiency in transaction processing and supporting query requirements *i.e.,* they cannot be further optimized for the applications such as OLAP, DSS and data mining.	A data warehouse is generally optimized to access from a decision maker's needs. These are designed specifically to support efficient extraction, processing and presentation for analytical and decision- making purpose.
3.	Database are generally small as compared to Data warehouses and contain data from single source.	Data warehouses generally contain very large amounts of data from multiple sources that may include databases from different data models and sometimes files acquired from independent systems and platforms.
4.	Multi databases provide access to disjoint and usually heterogeneous data bases and are volatile.	Data warehouse is generally a store of integrated data from multiple sources, processed for storage in a multi- dimensional model and nonvolatile.
5.	They do not support time series and trend analysis.	Data warehouses support time-series and trend analysis, both of which require more historical data.

6.	In databases, transactions are the unit and are the agent of change to the database.	The information in the data warehouses is much more course–grained and is refreshed according to a careful choice of incremental refresh policy.
7.	Data in the databases can be changed regularly.	Data can only be added. Once data are stored, no changes are allowed.
8.	Data represent current view.	Data are historic in nature.
9.	Same data can have different representations or meanings.	It provides a unified view of all data elements with a common definition and representation.
10.	Data are stored with a functional or process orientation.	Data are stored with a subject orientation.

3. Data Warehouse Architecture

The hardware, software and data resources required to construct the data warehouse depends upon the organization that wants to construct it. The needs and resources available forces the decisions of the organization regarding the architecture of a particular data warehouse. There are many phases, that are common to all the data warehouses regardless of the organization or the design selected. The most common phases are acquisition of data, storage of data and data access. The general architecture of a data warehouse is shown in Figure 1.

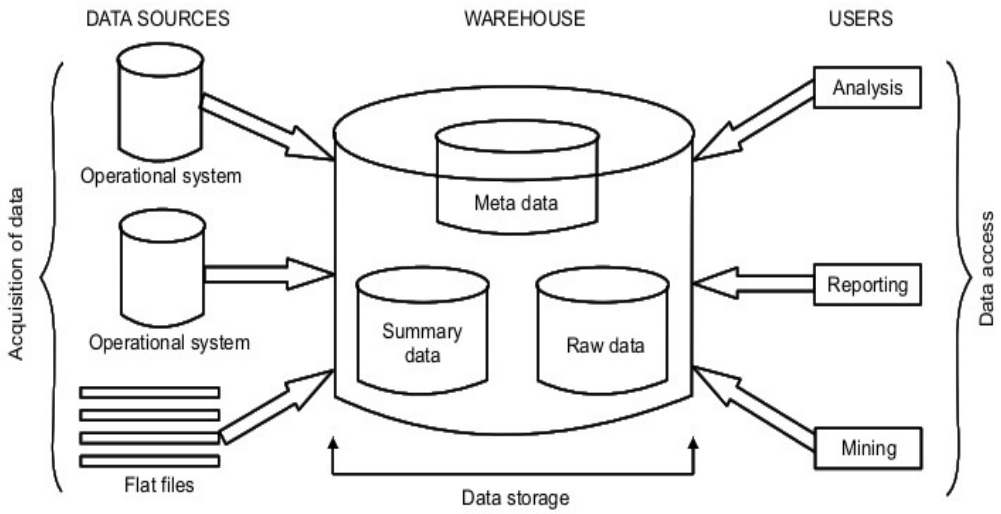

FIGURE 1. Data warehouse architecture.

Acquisition of Data: All the data warehouses must have a source from where the data is acquired. Most of the data in the data warehouse is derived from the operational data of the organization. The required data is extracted, filtered, translated and integrated into the data storage environment.

Storage of Data: The large amounts of operational data that is historical in nature are defined, indexed and then partitioned to allow for economic and efficient access.

Data Access: A number of data mining applications allow many users throughout the organization to retrieve, analyze, query and generate reports. The ability to access data is fundamental to the concept of data warehouse in the organization.

4. Data Warehouse Components

There are mainly six components of a data warehouse. These are as follows:

(i) Summarized data

(ii) Operational data-store

(iii) Integration/Transformation programs

(iv) Detailed data

(v) Metadata

(vi) Archives

1) **Summarized data:** The raw data generated by a transaction-processing system may be too large to store online. However, many queries can be answered by just maintaining the summary data obtained by aggregation on a relation, rather than maintain the entire relation. Summary data is classified into two categories—Lightly summarized and Highly summarized.
    - ***Lightly summarized data*:** This represents data distilled from current detailed data. It is summarized according to some unit of time and always resides on disk.
    - ***Highly summarized data*:** This represents data distilled from lightly summarized data. It is more compact and easily accessible and resides on disk.
2) **Operational data store :** Operational databases are the source data for the data warehouse. Operational data store is a repository of operationaldata.
3) **Integration/transformation programs :** The integration and transformation programs convert the operational data that is applications specific into enterprise data. The major functions performed by these programs are as follows:
    - Reformatting, re-evaluation or changing key structures.
    - Adding time elements.
    - Default values identification.
    - Providing logic to choose between multiple data sources.
    - Summarizing, tallying and merging data from multiple sources.

These programs are modified when operational or data warehouse environments change to reflect the changes.

4) **Detailed data :** Detailed data is of two types—Older detail data and current detail data. The older detail data represent data that is not very recent, may be as old as ten years or longer. It is voluminous and most frequently stored on mass storage such as tape. The current detail data represent data of a recent nature and always has a shorter time horizon than older detail data. It can be voluminous; it is almost always stored on disk to permit faster access.

5) **Meta data :** Meta data is data about data. Meta data for data warehouse users are part of the data warehouse itself and controls access and analysis of the data warehouse contents. The meta data repository is a key data warehouse component. It contains both technical and business meta data. The technical meta data cover details about acquisition, processing, storage structure, data descriptions, warehouse operations and maintenance and access support functionality. The business meta data covers the relevant business rules and organizational details supporting the warehouse.

6) **Archives :** These contain old or historical data of significant interest and have value to the enterprise. It is generally used for forecasting and trend analysis, thus, these archives store old data and the meta data that describe the characteristics of the old data.

5. Advantages of Data Warehouse:

The data warehouse has many advantages for an enterprise. The most important one are as follows.

1) **Effective decision making :** The major benefit of a data warehouse is its ability to analyze and execute business decisions based on data from multiple sources. By using data warehouse, one can look at past trends and may be do some predictions of what is going to happen in the future.

2) **Increases the productivity of business analysts :** The data warehouse can provide analysts with pre-calculated reports and graphs, that increase the productivity of business analysts.

3) An enterprise can maintain better customer relationships by correlating all customer data through a single data warehouse.

4) It provides supplementing disaster recovery plans with another data backup source.

5) **Business and information re-engineering**: By knowing what information is important to the enterprise, that is possible by using data warehousing, the re-engineering efforts become more directional and have priorities. Also the data warehouse development is the effective first step in re-engineering the enterprise's legacy system.

6. Disadvantages/Limitations of Data Warehouse

1. The data warehouse have some limitations, these are as follows:
2. The data warehouse is very expensive solution and generally found in large firms.
3. Performance tuning is hard due to very large size of the data warehouse.
4. The cost of maintaining the data warehouse is very high.
5. A data warehouse has a high demand of various resources.
6. Scalability can be a problem with the data warehouse.
7. Complexity of integration in data warehouse.
8. Data warehouse is query intensive.

Assignment:

Que-1. What are data warehouses? Write down the Data Warehouse Components.

Que-2. Differentiates between Database and Data Warehouse.

Que-3. Explain Data Warehouse Architecture with its advantages and Disadvantages.